



OXFORD

Ecological Statistics

CONTEMPORARY THEORY AND APPLICATION

Edited by GORDON A. FOX,
SIMONETA NEGRETE-YANKELEVICH,
AND VINICIO J. SOSA

Ecological Statistics: Contemporary Theory and Application

Edited By

GORDON A. FOX

University of South Florida

SIMONETA NEGRETE-YANKELEVICH

Instituto de Ecología A. C.

VINICIO J. SOSA

Instituto de Ecología A. C.



OXFORD
UNIVERSITY PRESS

Ecological Statistics: Contemporary Theory and Application. First Edition. Edited by Gordon A. Fox, Simoneta Negrete-Yankelevich, and Vinicio J. Sosa. © Oxford University Press 2015.
Published in 2015 by Oxford University Press.

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Oxford University Press 2015

The moral rights of the authors have been asserted

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2014956959

ISBN 978-0-19-967254-7 (hbk.)

ISBN 978-0-19-967255-4 (pbk.)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Ecological Statistics: Contemporary Theory and Application

Dedication

Gordon A. Fox – To Kathy, as always.

Simoneta Negrete-Yankelevich – A Laila y Aurelio, con amor infinito.

Vinicio J. Sosa – To Gaby, Eras and Meli.

Acknowledgments

The contributors did much more than write their chapters; they provided invaluable help in critiquing other chapters and in helping to think out many questions about the book as a whole. We would like to especially thank Ben Bolker for his thinking on many of these questions. Graciela Sánchez Ríos provided much-needed help with the bibliography. Fox was supported by grant number DEB-1120330 from the U.S. National Science Foundation. The Instituto de Ecología A.C. (INECOL) encouraged this project from beginning to end, and gracefully allocated needed funding (through the *Programa de Fomento a las Publicaciones de Alto Impacto/Avances Conceptuales y Patentes 2012*) to allow several crucial work meetings of the editors; without this help this book would probably not have seen the light of day.

List of contributors

Benjamin M. Bolker

Departments of Mathematics & Statistics and Biology
McMaster University
1280 Main Street West
Hamilton, Ontario L8S 4K1
Canada
bolker@mcmaster.ca

Yvonne M. Buckley

School of Natural Sciences
Trinity College , University of Dublin
Dublin 2
Ireland
buckleyy@tcd.ie
and
The University of Queensland
School of Biological Sciences
Queensland 4072
Australia

Gordon A. Fox

Department of Integrative Biology (SCA 110)
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620
USA
gfox@usf.edu

James B. Grace

US Geological Survey
700 Cajundome Blvd.
Lafayette, LA 70506
USA
gracej@usgs.gov

Jessica Gurevitch

Department of Ecology and Evolution
Stony Brook University
Stony Brook, NY 11794-5245
USA
Jessica.Gurevitch@stonybrook.edu

Bruce E. Kendall

Bren School of Environmental Science & Management
University of California, Santa Barbara
Santa Barbara CA 93106-5131
USA
kendall@bren.ucsb.edu

Marc J. Lajeunesse

Department of Integrative Biology (SCA 110)
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620
USA
lajeunesse@usf.edu

Michael A. McCarthy

School of BioSciences
The University of Melbourne
Parkville VIC 3010
Australia
mamcca@unimelb.edu.au

Earl D. McCoy

Department of Integrative Biology (SCA 110)
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620
USA
edm@mail.usf.edu

Shinichi Nakagawa

Department of Zoology
University of Otago
340 Great King Street
P.O. Box 56
Dunedin
New Zealand
shinichi.nakagawa@otago.ac.nz

and

School of Biological, Earth and Environmental Sciences
University of New South Wales
Sydney
NSW 2052
Australia

Simoneta Negrete-Yankelevich

Instituto de Ecología A. C. (INECOL)
Carretera Antigua a Coatepec 351
El Haya Xalapa 91070

Veracruz
México
simoneta.negrete@inecol.mx

Jonathan R. Rhodes

The University of Queensland
School of Geography, Planning, and Environmental Management
Brisbane
Queensland 4072
Australia
jrhodes@uq.edu.au

Shane A. Richards

School of Biological & Biomedical Sciences
Durham University
South Road
Durham, DH1 3LE
UK
s.a.richards@durham.ac.uk

Samuel M. Scheiner

Division of Environmental Biology
National Science Foundation
Arlington, VA 22230
USA
sscheine@nsf.gov

Donald R. Schoolmaster Jr.

U. S. Geological Survey
700 Cajundome Blvd.
Lafayette, LA 70506
USA
schoolmasterd@usgs.gov

Vinicio J. Sosa

Instituto de Ecología A. C. (INECOL)
Carretera Antigua a Coatepec 351
El Haya Xalapa 91070
Veracruz
México
vinicio.sosa@inecol.mx

Contents

<i>List of contributors</i>	xiii
Introduction <i>Vinicio J. Sosa, Simoneta Negrete-Yankelevich, and Gordon A. Fox</i>	1
Why another book on statistics for ecologists?	1
Relating ecological questions to statistics	5
A conceptual foundation: the statistical linear model	7
What we need readers to know	12
How to get the most out of this book	13
1 Approaches to statistical inference <i>Michael A. McCarthy</i>	15
1.1 Introduction to statistical inference	15
1.2 A short overview of some probability and sampling theory	16
1.3 Approaches to statistical inference	19
1.3.1 Sample statistics and confidence intervals	20
1.3.2 Null hypothesis significance testing	21
1.3.3 Likelihood	27
1.3.4 Information-theoretic methods	30
1.3.5 Bayesian methods	33
1.3.6 Non-parametric methods	39
1.4 Appropriate use of statistical methods	39
2 Having the right stuff: the effects of data constraints on ecological data analysis <i>Earl D. McCoy</i>	44
2.1 Introduction to data constraints	44
2.2 Ecological data constraints	45
2.2.1 Values and biases	45
2.2.2 Biased behaviors in ecological research	47
2.3 Potential effects of ecological data constraints	48
2.3.1 Methodological underdetermination and cognitive biases	48
2.3.2 Cognitive biases in ecological research?	49
2.4 Ecological complexity, data constraints, flawed conclusions	50
2.4.1 Patterns and processes at different scales	51
2.4.2 Discrete and continuous patterns and processes	52
2.4.3 Patterns and processes at different hierarchical levels	54
2.5 Conclusions and suggestions	56
3 Likelihood and model selection <i>Shane A. Richards</i>	58
3.1 Introduction to likelihood and model selection	58
3.2 Likelihood functions	59

3.2.1	Incorporating mechanism into models	61
3.2.2	Random effects	63
3.3	Multiple hypotheses	65
3.3.1	Approaches to model selection	67
3.3.2	Null hypothesis testing	68
3.3.3	An information-theoretic approach	70
3.3.4	Using AIC to select models	73
3.3.5	Extending the AIC approach	74
3.3.6	A worked example	76
3.4	Discussion	78
4	Missing data: mechanisms, methods, and messages	
	<i>Shinichi Nakagawa</i>	81
4.1	Introduction to dealing with missing data	81
4.2	Mechanisms of missing data	83
4.2.1	Missing data theory, mechanisms, and patterns	83
4.2.2	Informal definitions of missing data mechanisms	83
4.2.3	Formal definitions of missing data mechanisms	84
4.2.4	Consequences of missing data mechanisms: an example	86
4.3	Diagnostics and prevention	88
4.3.1	Diagnosing missing data mechanisms	88
4.3.2	How to prevent MNAR missingness	90
4.4	Methods for missing data	92
4.4.1	Data deletion, imputation, and augmentation	92
4.4.2	Data deletion	92
4.4.3	Single imputation	92
4.4.4	Multiple imputation techniques	94
4.4.5	Multiple imputation steps	95
4.4.6	Multiple imputation with multilevel data	98
4.4.7	Data augmentation	101
4.4.8	Non-ignorable missing data and sensitivity analysis	101
4.5	Discussion	102
4.5.1	Practical issues	102
4.5.2	Reporting guidelines	103
4.5.3	Missing data in other contexts	104
4.5.4	Final messages	105
5	What you don't know can hurt you: censored and truncated data in ecological research	
	<i>Gordon A. Fox</i>	106
5.1	Censored data	106
5.1.1	Basic concepts	106
5.1.2	Some common methods you should not use	107
5.1.3	Types of censored data	109
5.1.4	Censoring in study designs	111
5.1.5	Format of data	113
5.1.6	Estimating means with censored data	113
5.1.7	Regression for censored data	116

5.2	Truncated data	124
5.2.1	Introduction to truncated data	124
5.2.2	Sweeping the issue under the rug	125
5.2.3	Estimation	125
5.2.4	Regression for truncated data	127
5.3	Discussion	129
6	Generalized linear models <i>Yvonne M. Buckley</i>	131
6.1	Introduction to generalized linear models	131
6.2	Structure of a GLM	135
6.2.1	The linear predictor	135
6.2.2	The error structure	136
6.2.3	The link function	136
6.3	Which error distribution and link function are suitable for my data?	137
6.3.1	Binomial distribution	138
6.3.2	Poisson distribution	141
6.3.3	Overdispersion	143
6.4	Model fit and inference	145
6.5	Computational methods and convergence	146
6.6	Discussion	147
7	A statistical symphony: instrumental variables reveal causality and control measurement error <i>Bruce E. Kendall</i>	149
7.1	Introduction to instrumental variables	149
7.2	Endogeneity and its consequences	151
7.2.1	Sources of endogeneity	152
7.2.2	Effects of endogeneity propagate to other variables	154
7.3	The solution: instrumental variable regression	154
7.3.1	Simultaneous equation models	158
7.4	Life-history trade-offs in Florida scrub-jays	158
7.5	Other issues with instrumental variable regression	161
7.6	Deciding to use instrumental variable regression	163
7.7	Choosing instrumental variables	165
7.8	Conclusion	167
8	Structural equation modeling: building and evaluating causal models <i>James B. Grace, Samuel M. Scheiner, and Donald R. Schoolmaster, Jr.</i>	168
8.1	Introduction to causal hypotheses	168
8.1.1	The need for SEM	168
8.1.2	An ecological example	169
8.1.3	A structural equation modeling perspective	171
8.2	Background to structural equation modeling	173
8.2.1	Causal modeling and causal hypotheses	173

8.2.2	Mediators, indirect effects, and conditional independence	174
8.2.3	A key causal assumption: lack of confounding	175
8.2.4	Statistical specifications	175
8.2.5	Estimation options: global and local approaches	176
8.2.6	Model evaluation, comparison, and selection	178
8.3	Illustration of structural equation modeling	179
8.3.1	Overview of the modeling process	179
8.3.2	Conceptual models and causal diagrams	180
8.3.3	Classic global-estimation modeling	181
8.3.4	A graph-theoretic approach using local-estimation methods	186
8.3.5	Making informed choices about model form and estimation method	190
8.3.6	Computing queries and making interpretations	193
8.3.7	Reporting results	196
8.4	Discussion	197
9	Research synthesis methods in ecology <i>Jessica Gurevitch</i> <i>and Shinichi Nakagawa</i>	200
9.1	Introduction to research synthesis	200
9.1.1	Generalizing from results	200
9.1.2	What is research synthesis?	201
9.1.3	What have ecologists investigated using research syntheses?	201
9.1.4	Introduction to worked examples	202
9.2	Systematic reviews: making reviewing a scientific process	203
9.2.1	Defining a research question	204
9.2.2	Identifying and selecting papers	204
9.3	Initial steps for meta-analysis in ecology	204
9.3.1	What not to do	205
9.3.2	Data: What do you need, and how do you get it?	205
9.3.3	Software for meta-analysis	207
9.3.4	Exploratory data analysis	207
9.4	Conceptual and computational tools for meta-analysis	210
9.4.1	Effect size metrics	210
9.4.2	Fixed, random and mixed models	210
9.4.3	Heterogeneity	211
9.4.4	Meta-regression	213
9.4.5	Statistical inference	213
9.5	Applying our tools: statistical analysis of data	214
9.5.1	Plant responses to elevated CO ₂	214
9.5.2	Plant growth responses to ectomycorrhizal (ECM) interactions	220
9.5.3	Is there publication bias, and how much does it affect the results?	221
9.5.4	Other sensitivity analyses	222
9.5.5	Reporting results of a meta-analysis	223
9.6	Discussion	224
9.6.1	Objections to meta-analysis	224
9.6.2	Limitations to current practice in ecological meta-analysis	226
9.6.3	More advanced issues and approaches	226

10 Spatial variation and linear modeling of ecological data	
<i>Simoneta Negrete-Yankelevich and Gordon A. Fox</i>	228
10.1 Introduction to spatial variation in ecological data	228
10.2 Background	232
10.2.1 Spatially explicit data	232
10.2.2 Spatial structure	232
10.2.3 Scales of ecological processes and scales of studies	236
10.3 Case study: spatial structure of soil properties in a <i>milpa</i> plot	237
10.4 Spatial exploratory data analysis	238
10.5 Measures and models of spatial autocorrelation	239
10.5.1 Moran's I and correlograms	240
10.5.2 Semi-variance and the variogram	242
10.6 Adding spatial structures to linear models	246
10.6.1 Generalized least squares models	247
10.6.2 Spatial autoregressive models	250
10.7 Discussion	259
11 Statistical approaches to the problem of phylogenetically correlated data	
<i>Marc J. Lajeunesse and Gordon A. Fox</i>	261
11.1 Introduction to phylogenetically correlated data	261
11.2 Statistical assumptions and the comparative phylogenetic method	262
11.2.1 The assumptions of conventional linear regression	263
11.2.2 The assumption of independence and phylogenetic correlations	265
11.2.3 What are phylogenetic correlations and how do they affect data?	266
11.2.4 Why are phylogenetic correlations important for regression?	272
11.2.5 The assumption of homoscedasticity and evolutionary models	278
11.2.6 What happens when the incorrect model of evolution is assumed?	280
11.3 Establishing confidence with the comparative phylogenetic method	281
11.4 Conclusions	283
12 Mixture models for overdispersed data	
<i>Jonathan R. Rhodes</i>	284
12.1 Introduction to mixture models for overdispersed data	284
12.2 Overdispersion	286
12.2.1 What is overdispersion and what causes it?	286
12.2.2 Detecting overdispersion	288
12.3 Mixture models	289
12.3.1 What is a mixture model?	289
12.3.2 Mixture models used in ecology	292
12.4 Empirical examples	293
12.4.1 Using binomial mixtures to model dung decay	293
12.4.2 Using Poisson mixtures to model lemur abundance	299
12.5 Discussion	306
13 Linear and generalized linear mixed models	
<i>Benjamin M. Bolker</i>	309
13.1 Introduction to generalized linear mixed models	309
13.2 Running examples	310

13.3	Concepts	311
13.3.1	Model definition	311
13.3.2	Conditional, marginal, and restricted likelihood	319
13.4	Setting up a GLMM: practical considerations	322
13.4.1	Response distribution	322
13.4.2	Link function	323
13.4.3	Number and type of random effects	323
13.5	Estimation	323
13.5.1	Avoiding mixed models	324
13.5.2	Method of moments	324
13.5.3	Deterministic/frequentist algorithms	324
13.5.4	Stochastic/Bayesian algorithms	325
13.5.5	Model diagnostics and troubleshooting	326
13.5.6	Examples	327
13.6	Inference	328
13.6.1	Approximations for inference	328
13.6.2	Methods of inference	329
13.6.3	Reporting the GLMM results	331
13.7	Conclusions	333
	<i>Appendix</i>	335
	<i>Glossary</i>	345
	<i>References</i>	354
	<i>Index</i>	379

Introduction

*Vinicio J. Sosa, Simoneta Negrete-Yankelevich,
and Gordon A. Fox*

Why another book on statistics for ecologists?

This is a fair question, given the number of available volumes on the subject. The reason is deceptively simple: our use and understanding of statistics has changed substantially over the last decade or so. Many contemporary papers in major ecological journals use statistical techniques that were little known (or not yet invented) a decade or two ago. This book aims at synthesizing a number of the major changes in our understanding and practice of ecological statistics.

There are several reasons for this change in statistical practice. The most obvious cause is the continued growth of computing power and the availability of software that can make use of that power (including, but by no means restricted to, the R language). Certainly, the notebook and desktop computers of today are vastly more powerful than the mainframe computers that many ecologists (still alive and working today) once had to use. Both hardware and software can still impose limits on the questions we ask, but the constraints are less severe than in the past.

The ability to ask new questions, together with a growing body of practical experience and a growing cadre of ecological statisticians, has led to an increased level of statistical sophistication among ecologists. Today, many ecologists recognize that the questions we ask should be dictated by the scientific questions we would like to address, and not by the limitations of our statistical toolkit. You may be surprised to hear that this has ever been an issue, but letting our statistical toolkit determine the questions we address was a dominant practice in the past and is still quite common. However, increasingly today we see ecologists adapting procedures from other disciplines, or developing their own, to answer the questions that arise from their research. This change in statistical practice is what we mean by “deceptively simple” in the first paragraph: the difference between ecologists’ statistical practice today and a decade or two ago is not just that we can compute quantities more quickly, or crunch more (complex) data. We are using our data to consider problems that are more complex. For example, a growing number of studies use statistical methods to estimate parameters (say, the probability that the seed of an invasive pest will disperse X meters) for use in models that consider questions like rates of population growth or spread, risks of extinction, or changes to species’ ranges; fundamental questions, but ones that were previously divorced from statistics. Meaningful estimates of these quantities require careful choice of statistical approaches, and sometimes these approaches cannot be

limited to the contents of traditional statistics courses. This is of course only a point in a continuum; future techniques will continue to extend our repertoire of tractable questions and new books like this will continue to appear.

There is nothing wrong with using basic or old statistical techniques. Techniques like linear regression and analysis of variance (ANOVA) are powerful, and we continue to use them. But using techniques because we know them (rather than because they are appropriate) amounts to fitting things into a Procrustean bed—it does not necessarily ask the question we want to ask. We encountered recently a small but illustrative example in one of our labs: identifying environmental characteristics predicting presence of a lily, *Lilium catesbaei* (Sommers et al. 2011). It seemed reasonable to approach this problem with logistic regression (GLM with a binomial link; chapter 6), using site characteristics as the predictors and probability of presence/absence as the outcome. In reviewing literature on prediction of site occupancy, we found that a very large fraction of studies used a very different approach: ANOVA to compare the mean site characteristics of occupied with unoccupied sites. These might seem like comparable approaches, but they are quite different: logistic regression models probability of occupancy as a function of site characteristics, while ANOVA considers occupancy to be like an experimental treatment that somehow causes site characteristics! Yet many studies had used just this approach. To explore the problem, we analyzed the data using both approaches. The set of explanatory variables that we found predicted lily presence (using logistic regression) was not the same as the set of predictors for which occupied and unoccupied sites differed significantly (using ANOVA). The difference is not because the two approaches differ in power, or because we strongly violated underlying assumptions using one of the methods; the different results occur because the questions asked by the two approaches are quite different. This underlines a point that is often not obvious to beginners: the same data processed with different methods leads to different answers. By choosing a statistical method because it is convenient, we run the risk of answering questions we do not intend to ask. Worse still, we may not even realize that we have answered the wrong question.

The idea for this book emerged during a couple of occasions on which Fox came to Mexico to teach a survival module in the Sosa–Negrete statistics course for ecology graduate students. Dinner conversations often converged on the conclusion that, despite considerable efforts, learning statistics continues to be boring for many ecologists and more often than not, it feels a bit like having dental work done: frightening and painful but necessary for survival.

However, nothing could be further from the truth. Statistics is at the core of our science, because it provides us with tools that help us interpret our complex (and noisy) picture of the natural world (figure I.1). Ecologists today are leading in the development of a number of areas of statistics, and potentially we have a lot more to contribute. Many techniques used by ecologists are thoughtful, efficient, powerful, and diverse. For young ecologists to be able to keep up with this phenomenal advance, old ways of teaching statistics (based on memorizing which ready-made test to use for each data type) no longer suffice; ecologists today need to learn concepts enabling them to understand overarching themes. This is especially clear in the contribution that ecologists and ecological problems have made to the development of roll-your-own models (Hilborn and Mangel, 1997; Bolker, 2008).

The chapters of this book are by experienced ecologists who are actively working to upgrade ecologists' statistical toolkit. This upgrade involves developing models and statistical techniques, as well as testing the utility, usability, and power of these techniques in real ecological problems. Some of the techniques highlighted in the book are not new, but are underused in ecology, and can be a great aid in data analysis.